Dynamic analysis of discrete dependent variables: a didactic essay

Carroll, Glenn R.

Glenn R. Carroll

Dynamic Analysis of Discrete Dependent Variables: A Didactic Essay *

Zentrum für Umfragen, Methoden und Analysen e. V., Mannheim
ZUMA-Bericht Nr. 82/08
Mai 1982

* This paper was written during my stay as a Guest Professor at ZUMA, Mannheim. It is intended as a didactic paper making no original methodological contributions; indeed. I have borrowed heavily from the work of James Coleman, Michael Hannan and Nancy Tuma. The support of my colleagues and the staff at ZUMA is gratefully acknowledged.

1. Introduction

Social scientists frequently study phenomena which occur in time as discrete or qualitative changes in one or more of the characteristics of a social unit. Commonly - as in job shifts and organizational mergers - these changes correspond to the intuitive concept of a social event. But often the qualitative change of interest may not be readily visible; examples of less apparent transformations include the making of decisions, the act of conformity and alterations in emotional states.

When social scientist study qualitative changes, they routinely use static research methods which ignore the temporal dimensions of the research problem. The most widely used of these methods are probit and logit techniques, log-linear models for contingency tables, and regression analysis with dummy dependent variables. While these techniques each have specific merits, their application implies that the temporal process generating change has reached an equilibrium state. Since social scientists rarely support this assumption substantively, the use of these common procedures may defy the underlying intent of the investigator. This is turn may have serious consequences for interpretation of the findings. Moreover, for many research problems the interest lies as much in the process of change as in its outcome; here the usual techniques are entirely inappropriate as they provide no clue to the time path of change.

These problems can be readily overcome if the investigator abandons the equilibrium assumption and uses methods which explicitly incorporate time, known generally as dynamic methods. Unfortunately, the tools of dynamic analysis are less well-known to social scientists and the relevant discussions in the technical literature are often found to be inaccessible by nonspecialists. This situation is commonly taken to imply that the

methods of dynamic analysis are either too complex to learn, or too obscure to use in practical problems. I think this reasoning is incorrect on both counts. The methods of dynamic analysis for discrete variables are often less complex than their common counterparts, the probit and logit techniques. They are also often simple to apply and to interpret within a substantive framework.

In this paper, I hope to demonstrate the soundness of my positions on these points. I will try to show why it is important to use dynamic methods and how they can be interpreted intuitively. My purpose is entirely didactic and I will use many examples from actual analysis problems. Upon completion, the reader should have gained an elementary understanding of these models as well as the ability to apply them to basic research problems.

2. Conceptual Tools for Dynamic Analysis

My focus in this paper is soley with discrete dependent variables where the social units under study can be classified into one of several qualitative categories or states. Examples of variables with such qualitative state include the employment status of a person (possible states being 'employed' and 'unemployed'), the marital status of a person (states of 'single', 'married', 'divorced', and 'widowed'), and the structural form of the political regime of a country ('traditional monarchy', 'military rule', 'one party rule' and 'multiparty rule').

The purpose of a dynamic model is to describe mathematically the process by which the social units move from one qualitative state to the others. The description contains two essential ingredients. First, it summarizes the

relative frequencies of occupancy for the different states. Second, it provides information about the timings of moves between the different states.

I will discuss here a broad class of dynamic models known as finite-state, continuous-time stochastic models. The finite-state characteristic of the models indicates that a social unit can occupy only a countable number of possible states. These states are required to be mutually exclusive and exhaustive; that is, at every point in time a social unit must occupy one, and only one, qualitative state. The set of all possible states is known as the state space and is denoted by a state space variable. It is important to recognize that the state space is a construction of the investigator and depends upon the substantive questions of interest; it does not arise naturally from the data and different investigators will often construct different state spaces even when analyzing the same data. For example, data on career histories of individuals might be analyzed by one investigator as a two-state process of movement between the states of employment and unemployment. Another investigator might analyze the data as a four-state problem of movement between the states of full-time employment, part-time employment, unemployment and out of the labor force. The designation of state spaces depends upon substantive motivations.

The models I review here are depicted in continuous-time as opposed to discrete-time. This means, quite simply, that changes between states can occur at any point in time. Discrete-time models, in contrast, constrain the occurence of changes to only specified times, which are usually separated by intervals of equal duration. In the past, it has been thought that discrete-time models are preferable to continuous-time models because most longitudinal social science data are collected in panel form with waves

of equal duration. Such reasoning allows the data to dictate the form of the model; it is preferable to motivate the model substantively and, when possible, estimate its parameters from available data, whatever its form. Moreover, we are now learning that use of discrete-time models for processes which occur continuously in time can lead, under not unusual circumstances, to erroneous inferences.

The fundamental elements of a finite-state, continuous-time stochastic model are constructs from the theory of stochastic processes. I will explain each of these constructs and discuss how they can be estimated from social science data. For clarity of exposition, I will confine most of the discussion to a model with only two states and assume that the interest lies with the process of movement between states and the variables which affect this process. The concepts are easily generalized to more complex models and we examine several of these later in the paper.

The State Space

I have already discussed in general the state-space variable. Let me now define it more precisely as $Y(t)$, an integer-valued variable which indicates the state occupied by a unit at time t. In our two state example $Y(t)$ can take only two values, each an arbitrarily-chosen integer assigned exclusively to a particular state. It is customary to assign integers from one and count upwards; so in our example, $Y(t)$ may take only the values $Y(t) = 1$ or $Y(t) = 2$. If our substantive interest was to model the process of movement between the employment state of employed and unemployed, then we might assign the integer 1 to the employed state and 2 to the unemployed state. Then when $Y(t) = 1$ for a given individual, it signifies that this person is employed at time t.

State Probabilities

The state-space variable is merely an accounting variable that is used to signify which state is occupied or being referred to; it does not provide information about the process of change between states. For this purpose other constructs are used, the most basic being the state probability. State probabilities describe the probability of occupying each state, or if you prefer to think more concretely, the proportion of units occupying each state. Since we are concerned with a dynamic process, the state probabilities are functions of time and I shall use here $p_1(t)$ to indicate the probability of occupying state 1 at time t and $p_2(t)$ for state 2. More generally, we define the state probabilities for Y(t) as

$$p_Y(t) = Pr[Y(t) = y] \tag{1}$$

where pY(t) is the state probability and y is any of the possible realizations of Y(t). Since the state space must be mutually exclusive and exhaustive, it follows directly that

$$\sum_y p_Y(t) = 1 \tag{2}$$

for any t. This means simply that the sum of all state probabilities at any time point in the process will equal unity.

State probabilities can be easily estimated for any point in time. The unbiased estimate of $p_Y(t)$ is simply the ratio of cases occupying state Y(t) = y at t over the total number of cases. In our illustration then, $p_1(t) = \frac{\text{number units in state 1 at t}}{\text{total number of units at t}}$ and similarly for $p_2(t)$. Obviously, in many temporal processes, the values for $p_Y(t)$ will depend on the particular time t chosen to calculate the state probabilities. For example, suppose we are studying human mortality and our two states are alive and dead; time

is measured in years of age. The probability of occupying the death state will be very low for the young ages (small t) and will increase with age (large t). Thus the state probability for the alive state decreases with age and for the death state it increases. At some age, probably around 100 years, everyone in the study will have died and the state probability for death at this age and beyond will remain constant with a value of one; for the alive state it will remain at zero. When a process reaches such a point where the state probabilities no longer change with increases in time, the process is said to have reached equilibrium. It is important to recognize that equilibrium state probabilities need not leave all units in only one state as in our mortality example; indeed, some of the most interesting stochastic models display nonzero probabilities for all states when in equilibrium.

Transition Probabilities

Unlike state probabilities, which are unconditional and calculated for a single moment in time, transition probabilities are conditional upon the state occupied and calculated relative to two time points. Transition probabilities describe the probabilities of specific changes in the state space variable across two points in time. If we define two points in time as t and t+Δt such that Δt is always positive, then the state probability for a move from state j to state k is the probability of occupying state k at t+Δt given state j was occupied at t. For the two-state model, the four transition probabilities are

$$q_{11}(t,t+\Delta t) = \Pr[Y(t,t+\Delta t) = 1 \mid Y(t) = 1] \quad (3.1)$$

$$q_{12}(t,t+\Delta t) = \Pr[Y(t,t+\Delta t) = 2 \mid Y(t) = 1] \quad (3.2)$$

$$q_{21}(t,t+\Delta t) = \Pr[Y(t,t+\Delta t) = 1 \mid Y(t) = 2] \quad (3.3)$$

$$q_{22}(t,t+\Delta t) = \Pr[Y(t,t+\Delta t) = 2 \mid Y(t) = 2] \quad (3.4)$$

More generally, the transition probability can be defined as

$$q_{jk}(t,t+\Delta t) = Pr\left[\, Y(t+\Delta t) = k \mid Y(t) = j \,\right] \qquad (4)$$

where $Y(t) = j$ and $Y(t+\Delta t) = k$.

Transition probabilities are also easy to estimate empirically. The unbiased estimator of $q_{jk}(t,t+\Delta t)$ is simply the number of units which occupied state j at t and also occupied state k at $t+\Delta t$ divided by the total number of units occupying state j at time t. Obviously, the transition probabilities will vary depending upon the length and characteristics of the interval between t and $t+\Delta t$. Nonetheless, since only one state can be occupied at any single moment, the transition probabilities for any specific two times will always sum to one over all values of the origin state j.

Transition Rates

The workhorse of the models I will review here is the instantaneous transition rate. This construct - which I will often refer to as simply the rate - is defined as the transition probability over a unit of time where the unit is infinitesimal. More formally, the rate between two states j and k is defined as

$$r_{jk}(t) = \lim_{\Delta t \to 0} \frac{q_{jk}(t,t+\Delta t)}{\Delta t} \qquad (5)$$

where q_{jk} is the transition probability and Δt is assumed to be positive. Thus, the transition rate will always be nonnegative although it is unbounded above.

Transition rates are used as the focal points in dynamic analysis of discrete dependent variables because they uniquely determine the other constructs of the model (the reverse is not always true). However, unlike

probabilities, which have intuitive appeal, transition rates seem to many to be mysterious concepts, difficult to grasp concretely. This occurs, I think, because transition rates are unobservable and because we normally consider only rates calculated over fixed finite intervals such as one year. Unfortunate as this may be, it is important to develop some understanding of the instantaneous transition rate since it is the construct usually considered to be the dependent variable in a dynamic analysis of discrete variables.

How can transition rates be understood more intuitively? One good way is to think not about the values of the rate itself but instead to think about the values implied for the more intuitive concepts. As an example, let us examine the two-state model in detail. Since the non-movers are determined uniquely by the movers, we concentrate on the rates of movement across the two states 1 and 2. For simplicity, we assume that the rates are time-independent; that is

$$r_{12}(t) = r_{12} \tag{6.1}$$

$$r_{21}(t) = r_{21} \tag{6.2}$$

This assumption simply means that the transition rates are constant and do not vary with time.

Now suppose that time is measured in years and that the estimated transition rates are $\hat{r}_{12} = .5$ and $\hat{r}_{21} = 2.0$ (later we shall discuss how to estimate rates). As I have stated above, these values have no particular intuitive meaning for most but we can use them to estimate more intuitive parts of the model. In particular, for this model the state probabilities are determined by the rates according to the equations

$$p_1(t) = p_1(0)_e - (r_{12}+r_{21})t + \frac{r_{21}}{r_{12}+r_{21}}\left[1-e^{-(r_{12}+r_{21})t}\right] \quad (7.1)$$

$$p_2(t) = p_2(0)_e - (r_{12}+r_{21})t + \frac{r_{12}}{r_{12}+r_{21}}\left[1-e^{-(r_{12}+r_{21})t}\right] \quad (7.2)$$

where $p_j(0)$ is the initial proportion of units in state j at time 0, the beginning of the process. If we assume in our hypothetical case that these initial proportions are equally distributed (i.e., $p_1(0) = p_2(0) = 0.5$), then we can readily calculate the state probabilities for any time point. Table 1 gives some of these estimates for various times using the hypothetical rates. These estimates indicate the probability of being in either state at each point in the process; they are easy to grasp intuitively yet they are determined uniquely by the more mysterious transition rates.

Table 1 About Here

Table 1 also shows that the model we are examining eventually reaches a point where the state probabilities no longer change with time. This point is the equilibrium. Since it is stable, the proportion in state 1 will remain at .80 and in state 2 at .20. To calculate these equilibrium state probabilities directly for the two-state model, we can use the equations

$$p_1(\infty) = \frac{r_{21}}{r_{12}+r_{21}} \quad (8.1)$$

$$p_2(\infty) = \frac{r_{12}}{r_{12}+r_{21}} \quad (8.2)$$

where $p_j(\infty)$ signifies the equilibrium state probability. It is very important to recognize that although the state probabilities no longer change when equilibrium is reached, this does not imply that movement between states has ceased. Indeed, the rate of movement remains the same as before;

Table 1. Estimated State Probabilities for Hypothetical Transition Rates $r_{12} = .5$ and $r_{21} = 2.0$

t	$p_1(t)$	$p_2(t)$
0	.50	.50
0.5	.71	.29
1.0	.77	.23
2.0	.79	.21
5.0	.80	.20
10.0	.80	.20
100.0	.80	.20

it is just that the proportions in each state do not change as a result of this movement.

Transition probabilities between any two time points are also uniquely determined by the rates. For the two-state model, they can be calculated for the interval t, t+Δt by the equations

$$q_{12}(t,t+\Delta t) = \frac{r_{12}}{r_{12}+r_{21}}\left[1-e^{-(r_{12}+r_{21})\Delta t}\right] \qquad (9.1)$$

$$q_{21}(t,t+\Delta t) = \frac{r_{21}}{r_{12}+r_{21}}\left[1-e^{-(r_{12}+r_{21})\Delta t}\right] \qquad (9.2)$$

Considering our hypothetical example across an interval of one year (a single time unit; $\Delta t = 1$), we arrive at the transition probabilities $q_{12}(0,1) = .184$ and $q_{21}(0,1) = .734$. These estimates can be interpreted to mean that, given our rate of movement, 18.4% of those units in state 1 at time 0 moved to state 2 by time 1. Similarly, 73.4% of the occupants in state 2 at time 0 moved to state 1 by time 1. Since the model is time-independent, these transition probabilities hold true for any time interval of length 1 year, including within the equilibrium region.

Perhaps the most intuitive implications of a rate model are the expected durations in states that it generates. For the two-state, time-independent model, the average length of state occupancies are given by

$$E(u_1) = \frac{1}{r_{12}} \qquad (10.1)$$

for state 1, and

$$E(u_2) = \frac{1}{r_{21}} \qquad (10.2)$$

for state 2. Using the hypothetical data, this implies that average length of time spent in state 1 by its occupants is 2.0 years and it is 0.5 years for state 2. If one thinks of the change of state as an event,

then these figures can also be interpreted as the average waiting time until an event for state occupants. Either interpretation is intuitively appealling and the calculations are straightforward to make; many persons find thinking of rates in this manner as the easiest way to understand them.

Another intuitive way to understand a rate model is to consider a fixed length of time and to calculate the expected number of events that will occur in this period for a given rate. If we let $N_{jk}(0,t)$ represent the number of visits to state k within the period 0 to t by an occupant of state j at time 0, then we can find the average number of events in (0,t) for the two-state model by the equations

$$E\left[N_{jj}(0,t)\right] = \frac{r_{12}r_{21}t}{(r_{12}+r_{21})} + \frac{r_{12}r_{21}}{(r_{12}+r_{21})^2}\left[e^{-t(r_{12}+r_{21})} - 1\right] \tag{11.1}$$

$$E\left[N_{jk}(0,t)\right] = \frac{r_{12}r_{21}t}{(r_{12}+r_{21})} - \frac{r_{jk}^2}{(r_{12}+r_{21})^2}\left[e^{-t(r_{12}+r_{21})} - 1\right] \tag{11.2}$$

where j, k = 1, 2 and $j \neq k$. For our hypothetical rates considered over the time interval from 0 to 1, these equations predict

$$E\left[N_{11}(0,1)\right] = 0.25 \tag{12.1}$$

$$E\left[N_{12}(0,1)\right] = 0.44 \tag{12.2}$$

$$E\left[N_{21}(0,1)\right] = 0.25 \tag{12.3}$$

$$E\left[N_{22}(0,1)\right] = 0.99 \tag{12.4}$$

for the average number of events. $E\left[N_{22}(0,1)\right]$ can be interpreted as indicating that, on the average, each occupant of state 2 at time 0 will experience .99 additional visits to state 2 within the interval 0 to 1. Of course, to

experience additional visits to the state occupied initially, the state must be first left for another.

For research problems involving multiple kinds of events, it is often useful to decompose the transition rate into two conceptually distinct constructs. The first of these is the hazard function, which is simply the rate of leaving of a particular state irrespective of the destination state. So if we have k destination states, the hazard function is given by

$$h_j(t) = r_j(t) = \sum_{\substack{k \\ j \neq k}} r_{jk}(t) \quad . \tag{13}$$

In our two-state model, where only one destination is available to each origin state, the hazard function and the transition rates are identical.

The second element obtained in the decomposition of the transition rate is the conditional transition probability. This is defined as the probability of a move from state j to k, given that there has been a move. Its relationship to the rate and hazard function is

$$m_{jk}(t) = \frac{r_{jk}(t)}{h_j(t)} = \frac{r_{jk}(t)}{r_j(t)} \tag{14}$$

which shows that the conditional transition probability is the ratio of the probability of a move from j to k over the probability of leaving j. Re-arranging terms as

$$r_{jk}(t) = h_j(t) . m_{jk}(t) \tag{15}$$

shows the decomposition of the transition rate into a component for the rate of leaving state j (hazard function) and a component for the probability

of which state k will next be occupied. Decompositions of this variety are useful when movement between states is best viewed as a series of sequential acts: first leaving the state occupied and second deciding where to go. When these acts are seen as intermingled, it is best to work with the transition rate directly.

The rate models that we have considered so far are simple and often considered unrealistic. In research applications, two concerns with these models arise frequently. First, social processes are often thought to be time-dependent and to build a model reflecting this characteristic, the transition rate must be an explicit function of time. One example of such a specification is

$$r_{jk}(t) = e^{\alpha_{jk}} e^{\beta_{jk} t} \tag{16}$$

which is known as the Gompertz model. We shall examine it in greater detail later in the paper.

The second complexity that is often necessary to introduce in rate models is population heterogeneity. The models that we have considered to this point assume population homogeneity; that is, they assume that each unit in the study has the same rate and that its process of change is governed by the same parameters as all the other units. Often, however, the investigator will want to allow the rate of transition to depend on the specific characteristics of each individual social unit. For example, the rate of job change is believed to be a function of a person's age, workforce experience, education, income and sex. If we have data on these characteristics for each individual, then we can specify the rate as function

$$r_{jk}(t) = f(X_1, X_2, .., X_m) \tag{17}$$

of the exogenous variables X_1 to X_m. The simplest specification of this type is when the exogenous variables are linearly related to the rate, e.g.,

$$r_{jk}(t) = \alpha_o + \alpha_1 X_1(t) + \alpha_2 X_2(t) \tag{18}$$

where α_1 and α_2 are parameters measuring the effects of each variable on the rate. However, the linear model can sometimes lead to negative predicted rates, which are meaningless. For this reason, the log-linear specification

$$r_{jk}(t) = \exp(\alpha_c + \alpha_1 X_1(t) + \alpha_2 X_2(t)) \tag{19.1}$$

or

$$\ln \bar{r}_{jk}(t) = \alpha_c + \alpha_1 X_1(t) + \alpha_2 X_2(t) \tag{19.2}$$

is commonly used to introduce heterogeneity into the model.

When either time-dependence or population heterogeneity is introduced into the model, estimating rates and the parameters of the model is more difficult. In fact, for all but the simplest of these models, it is not possible to find explicit estimators as with the constant rate model. Instead, we are forced to rely on iterative techniques such as the maximum likelihood program RATE, developed by Nancy Tuma. Fortunately, however, these techniques can be accomplished quickly with the computer and the results have been shown to be highly reliable. Moreover, once given a predicted rate value for a sample unit, we can still employ the intuitive transformations used above to understand our estimated model. I shall illustrate this more completely later in the paper using models with both time-dependence and heterogeneity estimated with career history data on individuals.

Survivor Function

The final element of the model which we will examine here is the survivor function. The survivor function describes the probability of not having an event ('surviving') as a function of time. Suppose the event under study is death and time is measured in years of age from 0 to 100 years. A plot of the survivor function for this problem would

indicate the probability of surviving to each age (see Figure 1). The beginning probability for age zero would be one, because no one has died, and it would decline irreversibly with age (it is not possible to have a greater chance of living to 50 than 40). If everyone in the study dies before age 100, then the empirical probability of survival to this point would be zero. Alternatively, one can think of the survivor function as indicating the proportion surviving to each age from a group beginning at age zero.

Figure 1 About Here

More formally, we define the survivor function for state j as

$$G_j(u) = Pr\ (U \geq u) \tag{20}$$

where u is the realized duration in state j. Those familiar with probability theory will recognize the similarity of the survivor function to the cumulative distribution function, $F(t) = Pr\ [U \leq u]$. The relationship between the two is simply that $G(t) = 1-F(t)$.

The survivor function also has an exact relationship with the hazard function, or rate of leaving. There are two ways to show this relationship. First, the hazard can be written as a negative function of survivor function over changes in time

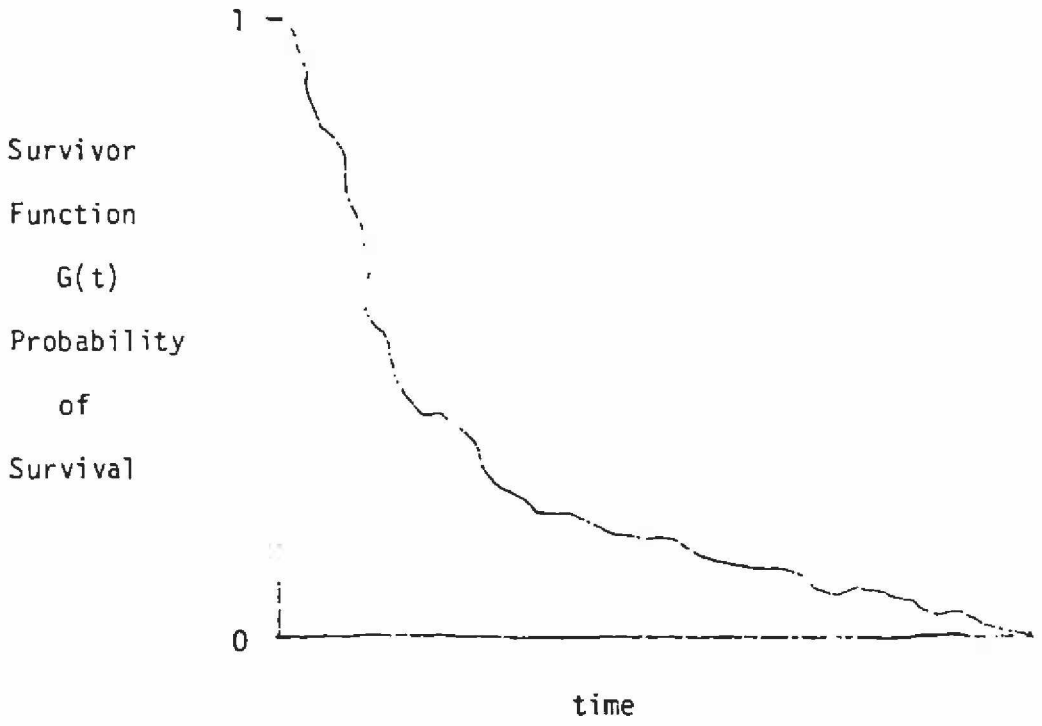
$$h_j(u) = \frac{-d \ln G_j(u)}{dt} \tag{21}$$

Second, the survivor function can be expressed in terms of the hazard rate integrated over time

$$G_j(t) = \exp -\int_0^t h_j(s)ds \tag{22}$$

Figure 1

Hypothetical Survivor Function



For the constant-rate, two-state model, this expression simplifies to

$$G_1(u) = e^{-h_1 u} \quad \text{and} \tag{23.1}$$

$$G_2(u) = e^{-h_2 u} \quad \text{, or equivalently,} \tag{23.2}$$

$$\ln G_1(u) = -h_1 u \tag{23.3}$$

$$\ln G_2(u) = -h_2 u \tag{23.4}$$

These equations imply that when the rate of leaving is time-independent, the rate has a linear relationship with the log of the survivor function. This relationship is critical for model search activities, because, although we must make parametric assumptions about the rate in order to estimate it, the same is not true for the survivor function. We can estimate it making no parametric assumptions (the technique is discussed in detail below and can be accomplished with SPSS). Therefore, a reasonable model search strategy is to estimate the survivor function empirically and to plot its logarithm against time. If the plot is linear, a time-independent model is appropriate; if it is non-linear, a time-dependent model may be helpful. In the linear case, the hazard rate can also be estimated from the plot by calculating the negative slope of the line across time.

We have now reviewed the fundamental elements of the model we will use in the applications below. The discussion will now become more concrete, focusing on data structures and specific models motivated by substantive concerns. For those who find it useful to consult the theoretical material during this discussion, the Appendices may be helpful. Appendix A

gives the definitions of the mathematical terms we have examined. Appendix B shows some useful relationships between these terms. Appendix C illustrates several functional forms which are commonly used in the specification of the transition rates.

3. Event-History Data Structures

Social science data come in a variety of forms and many of these can be used to estimate rate models. Often, however, certain implausible assumptions must be invoked to justify the estimation technique. For example, cross-sectional data can yield good estimates of the model provided that the process is assumed to be in equilibrium. Other types of data structures, such as panel data, require less stringent - although possibly equally incorrect - assumptions about the model.

Most of these estimation difficulties are overcome if the investigator has available event-history data; that is, data with information on the timings of events. Event-history data are the richest type of data for event-generating processes yet they remain unfamiliar to most social scientists. Consequently, it is worthwhile to review in detail the special considerations that must be taken in collecting, assembling and analyzing event-history data.

Figure 2 illustrates two hypothetical event-histories. The state space consists of two states and the second state is absorbing: units entering state 2 cannot return to state 1. The event of moving from state 1 to state 2 is, therefore, irreversible and might be used to model a process such as dying. Both individuals depicted in the figure begin the process at time t_0. Individual A moves from state 1 to state 2 at time t_1 and B makes the same move at t_2. The period of time each unit spends waiting for an event to occur is known as a spell or episode.

When, as with state 2, units cannot leave a state once it is entered, time in this state is not considered a spell. When, as with Individual B, the observation period ends before the unit has experienced an event, the spell is said to be censored. Thus, for the observation period in Figure 2, we can identify two spells: Individual A has a complete spell from t_0 to t_1, and Individual B has a censored spell from t_0 to t_1.

Figure 2 About Here

Data records for event-histories are organized around the spells of the process. For each spell, four pieces of critical information must be coded: the starting time of the spell (TS), the finishing time of the spell (TF), the starting state of the spell (SS) and the finishing state (SF). The top of Figure 3 shows how the event-histories in Figure 2 would be coded for use by the RATE program and most other programs. The two event histories contain two spells of data; each starting at TS = t_0 and SS = 1. The spell for individual A has a finishing state of 2 since we observe this move at t_1. Individual B's spell, however, is censored at t_1; this unit's finishing state is the same as its starting state.

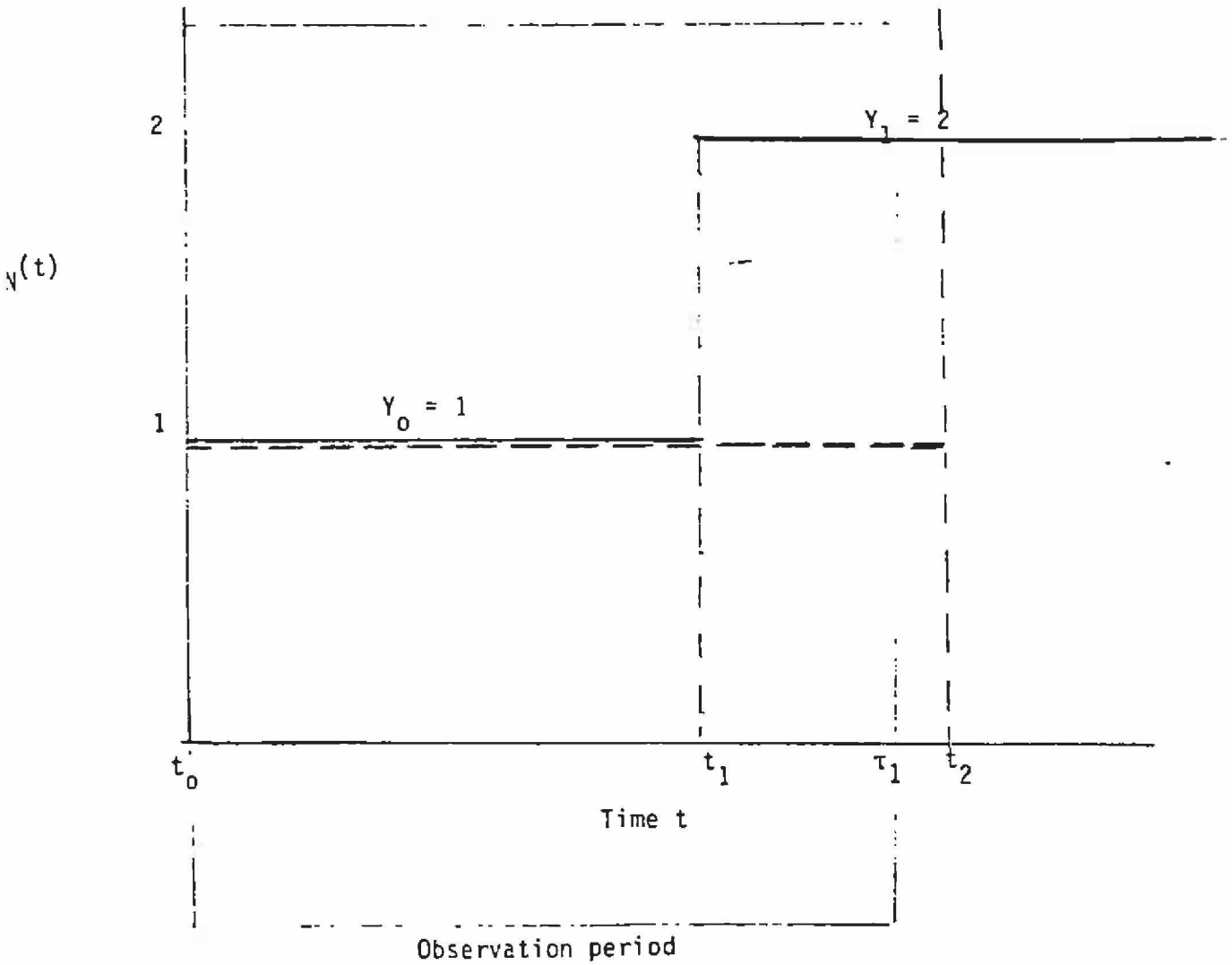
Figure 3 About Here

The data records also contain information on the characteristics of each individual which are thought to affect the rate, denoted in the figure by the X variables. These variables are assumed to be exogenous and are measured contemporaneously in synchronization with the time of the spell. Most frequently, the exogenous variables are measured at the starting time of each spell (see X_1 and X_2 in the figure); however,

Figure 2

Hypothetical Event-Histories

for Single Irreversible Events



Individual A

Individual B

Figure 3

Data Records for Hypothetical

Event-Histories in Figure 2

Alternative 1

I	TS	TF	SS	SF	$X_1(TS)$	$X_2(TS)$	$X_3(TF)...$
A	t_o	t_1	1	2	$X_{1A}(t_o)$	$X_{2A}(t_o)$	$X_{3A}(t_1)$
B	t_o	τ_1	1	1	$X_{1B}(t_o)$	$X_{2B}(t_o)$	$X_{3B}(\tau_1)$

Alternative 2: Assume TS = t_o = 0 and SS similiar for all units.

I	TF	SF	$X_1(TS)$	$X_2(TS)$	$X_3(TF)$
A	t_1	1	$X_{1A}(t_o)$	$X_{2A}(t_o)$	$X_{3A}(t_1)$
B	τ_1	0	$X_{1B}(t_o)$	$X_{2B}(t_o)$	$X_{3B}(\tau_1)$

this decision is partly substantive and the investigator may have strong reasons to measure some exogenous variables at the finishing time or some other point (see X_3 in the figure).

The bottom part of Figure 3 shows an alternative method of coding the event histories in Figure 2. The data here are still organized by spells and the exogenous variables remain unaffected. The difference between this coding scheme and the earlier one lies in the use of two assumptions which allow shortening of the data record. The first assumption is that TS = t_0 = 0 for all units. The second is that the starting state SS is the same for all units. The first assumption makes a difference only for time-dependent models that differentiate between calendar time and duration; the second assumption is already inherent in the process depicted in Figure 2. Use of these assumptions with the RATE program allows us to drop the TS and SS variables from the record. The only alteration required after this recission is to encode censored spells as SF = 0 (which is the usual convention). For single irreversible events, the integer associated with the destination state can also be changed for aesthetic appeal (see Individual A in Figure 3) but this is not really necessary.

In many research problems, the investigator will want to examine multiple destination states. For example, in my research on organizational mortality, I sometimes distinguish between death by disappearance and death by merger absorption. Figure 4 depicts hypothetical event-histories for such a process with multiple, irreversible destinations. Again individual A changes from state 1 to state 2 at time t_1. Individual B changes at time t_2, however, from state 1 to state 3. Since the starting

states for such a process will always be similar, these data can be coded as in Figure 5 provided we assume $t_o = 0$.

Figure 4 and Figure 5 About Here

Figure 5 is a straightforward translation of the graphic event histories into coded data: the finishing times and finishing states are simply transferred. With the TF and SF variables on these records, the RATE program could be used to estimate models with the rates $r_{12}(t)$ and $r_{13}(t)$ as the dependent variables. The problems with these variables, however, is that they are useful only for this purpose and none other. Often with data of this type, the investigator will want to examine first basic models for change to any destination state; then with this knowledge in hand, proceed to disaggregate the model into separate destination states. The first analysis could be accomplished, of course, by building a second data file organized similar to the records in the previous example and shown in Figure 3. A less cumbersome avenue, however, is simply to construct a second aggregated finishing state variable and include it on the original data records. The variable SF^* in Figure 5 shows an example of this procedure. These data records could now be used for two different analysis: one examining the rate of leaving the origin state; the other examining movement into state 2 and state 3 separately. For the first analysis SF^* and TF would be used; for the second, SF and TF.

Figure 6 depicts a more complex pair of event histories. In this characterization, there are three possible states and movement is possible from any one state to either of the two others. The data are again censored by the observation scheme but this time they are censored near the origin as well as near the end of the process. Censoring near the

Figure 4

Hypothetical Event-Histories with Multiple Irreversible Destination States

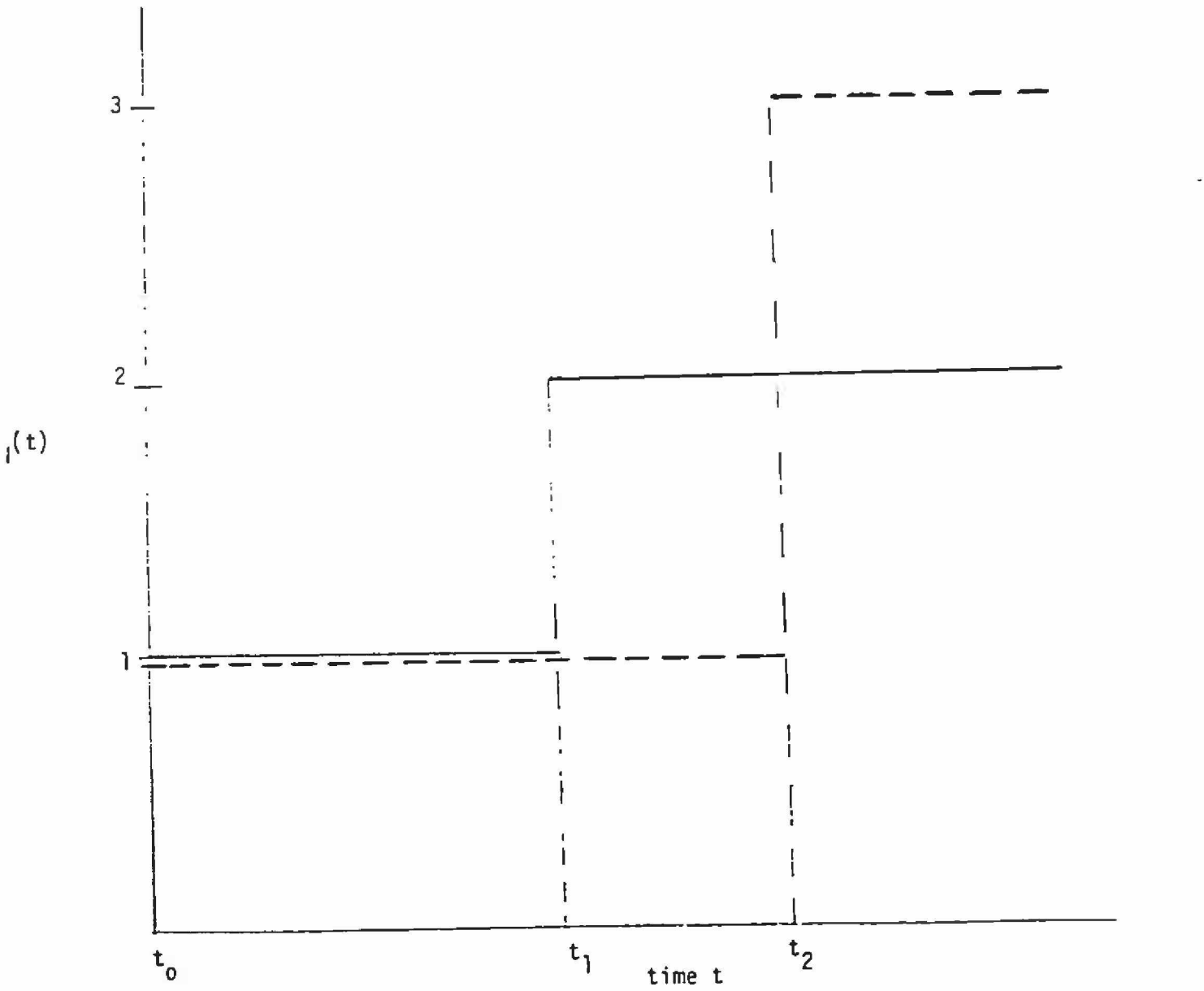


Figure 5

Data Records for

Event-Histories in Figure 4

I	TF	SF	$X_1(TS)$	$X_2(TS)$	$X_3(TF)$	SF*
A	t_1	2	$X_{1A}(t_o)$	$X_{2A}(t_o)$	$X_{3A}(t_1)$	1
B	t_2	3	$X_{1B}(t_o)$	$X_{2B}(t_o)$	$X_{3B}(t_2)$	1

origin is known as left-hand censoring and cannot always be treated adequately; it must be assumed either that $\tau_0 = t_0$ or that the period τ_0 to t_0 is irrelevant to the process. Censoring near the completion of the process is known as right-hand censoring and is accomodated in the finishing state variables.

Figure 6 About Here

Figure 7 shows an efficient method of coding the event-histories in Figure 6. Since Individual A has four observed spells, this unit receives four data records whereas the one spell for B translates into a single record. The exogenous variables are measured in coordination with the timing of spells so there are four measurements on each exogenous variable for A and only one for B. The variable N_I contains the sequence number of the spell for each individual and is often useful for data manipulation.

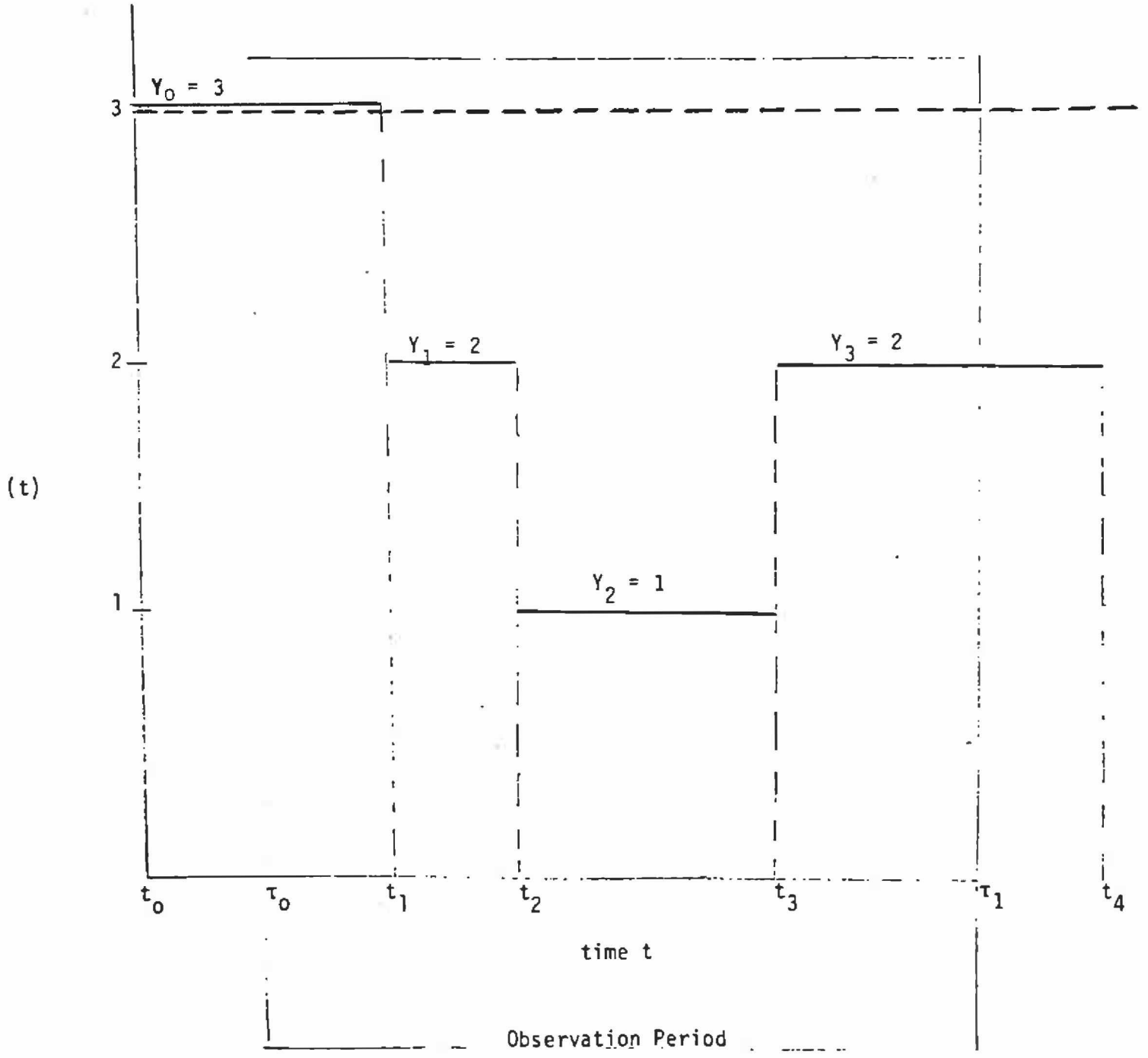
Figure 7 About Here

The data records in Figure 7 contain three time variables (TS, TF, TF^*) and three state variables (SS, SF, SF^*) which can be selected in a variety of ways depending on the problem to be studied. The simplest analysis, looking only at duration in the state and ignoring both origin state and destination state, would use TF^* and SF^*. This analysis could easily be extended to examine duration by each origin state to any destination (TF^*, SS, SF^*), or duration from any origin to each destination state (TF^*, SF). The full model, with six equations (one each for $r_{12}(t)$, $r_{13}(t)$, $r_{21}(t)$, $r_{23}(t)$, $r_{31}(t)$, $r_{32}(t)$), could also be examined by using TF^*, SS and SF. And, finally, if specific types of time-dependence were desired,

Figure 6

Hypothetical Event-Histories with

Multiple Origin and Destination States



Individual A ———

Individual B — — —

Adapted from Tuma (1979)

Figure 7

Data Records for

Event-Histories in Figure 6

I	TS	TF	TF*	SS	SF	SF*	N_I	$X_1(TS)$	$X_2(TS)$	$X_3(TF)$
A	τ_o	t_1	$t_1-\tau_o$	3	2	1	1	$X_{1A}(\tau_o)$	$X_{2A}(\tau_o)$	$X_{3A}(t_1)$
A	t_1	t_2	t_2-t_1	2	1	1	2	$X_{1A}(t_1)$	$X_{2A}(t_1)$	$X_{3A}(t_2)$
A	t_2	t_3	t_3-t_2	1	2	1	3	$X_{1A}(t_2)$	$X_{2A}(t_2)$	$X_{3A}(t_3)$
A	t_3	τ_1	τ_1-t_3	2	2	0	4	$X_{1A}(t_3)$	$X_{1A}(t_3)$	$X_{3A}(\tau_1)$
B	τ_o	τ_1	$\tau_1-\tau_o$	3	3	0	1	$X_{1B}(\tau_o)$	$X_{2B}(\tau_o)$	$X_{2B}(\tau_1)$

each of these specifications could be used with TS and TF instead of TF^*.

The data records in Figure 7 could be used to estimate eight fundamentally different models without any alteration, only selection by the analyst. With proper foresight, such a flexible event-history file can always be assembled provided that the different models to be explored involve only aggregation and disaggregation of the state space variable. This is often the case but certainly not always; many times the investigator will want to explore unique state space variables which recast the timings and number of spells altogether. For example, in analyzing career history data, one may wish to shift from an analysis of job changes to an analysis of movement between employment and unemployment. Since a single spell of employment may include many spells of jobs, the timings on the spells will change for the new problem; and most likely, so will the number of spells. When shifts this radical are desired, the investigator has no choice but to construct a new file.

4. Survival Analysis: Model Search Strategies

Event-history data can be used for description but most social scientists will wish a more analytic application. Occasionally, the analyst will have a developed theoretical model and wants to conduct a rigorous test of the empirical implications of this model. More commonly in the social sciences, however, the investigator has in mind some general theoretical notions about the relationships between several variables but does not have a precise specification of the form of the relationships. In these situations, the first step of the empirical analysis is an exploratory search for an appropriate model specification. Later analysis then proceeds within this framework.

When analyzing the rate of change in a discrete dependent variable, the exploratory analysis is conducted by examining empirical estimates of the survivor function G(t). A primary reason for concentrating on the survivor function at this stage is the availability of the nonparametric Kaplan-Meier estimator. This estimator allows us to make good estimates of any class of rate models without making any parametric or distributional assumptions. It is also easy to implement, available in software packages, and straightforward to interpret.

In this section of the paper, I demonstrate the use of the Kaplan-Meier estimator of the survivor function. I use career life-history data on 105 German individuals collected by Karl Ulrich Mayer. My focus is on the rate of job changes and the factors which affect it. The survival analysis reported here was conducted with the SURV routine of SPSS using a data structure similar to those discussed in the previous section.

The survivor function, you will recall, is a cumulative function which describes the probability of an event by any point in time. The Kaplan-Meier estimator assumes that event-history data are available and that the times of the observed events for all individuals can be ordered such that

$$t^*_{(1)} < t^*_{(2)} < \cdots < t^*_{(i)} \tag{24}$$

where $t^*_{(i)}$ are the times of the i-th observed event. Then if we let N represent the total number of units and $C(t_{(i)})$ represent the number of censored units with right-hand censoring times less than $t_{(i)}$, the Kaplan-Meier estimator of G(t) is

$$\hat{G}(t) = \prod_{t_{(i)}<t} \frac{N-i-C(t_{(i)})}{N-i-C(t_{(i)})+1} \tag{25}$$

which looks ominous due to notation. We can recast the estimator more intuitively if we introduce the concept of risk set. The risk set $R_{(i)}$ for the i-th event is the number of units still observed at the instant prior to $t^{*}_{(i)}$; in other words, the number of units "at risk" to experience the event. The risk set for the i-th event is the number of total units subtracting the number of units with previous events (i-1) and the number of previously censored units. Using the earlier notation, $R_{(i)} = N(i-1) - C(t_{(i)})$ and can be substituted in to yield

$$\hat{G}(t) = \prod_{t_{(i)}<t} \frac{R_{(i)}-1}{R_{(i)}} \qquad (26)$$

which shows the estimator to be a multiplicative function of the changing risk set with boundaries of zero and unity.

A simple hypothetical example will show the ease with which the survivor function can be estimated. Suppose we observe eight units with individual waiting times of 1.0, 3.0, 3.5^{+}, 4.2, 4.6^{+}, 4.8^{+}, 5.0, 6.8^{+} where + in the superscript denotes a censored observation. Since the times are already ordered, we need only apply the formula for the estimator to obtain an estimate of G(t) at each time of an observed event. This yields the estimates

$$\hat{G}(1.0) = \left(\frac{8-1}{8}\right) = 0.875$$

$$\hat{G}(3.0) = \left(\frac{8-1}{8}\right)\left(\frac{7-1}{7}\right) = 0.750$$

$$\hat{G}(4.2) = \left(\frac{8-1}{8}\right)\left(\frac{7-1}{7}\right)\left(\frac{5-1}{5}\right) = 0.600$$

$$\hat{G}(5.0) = \left(\frac{8-1}{8}\right)\left(\frac{7-1}{7}\right)\left(\frac{5-1}{5}\right)\left(\frac{2-1}{2}\right) = 0.300$$

and shows how the multiplicative form of the estimator leads to strictly decreasing values across time.

In large analysis problems, of course, estimating the survivor function can be an unwielding task and the computer must be used. Programming the Kaplan-Meier estimator is not a major task but there are several packaged routines readily available, including SURV in the widely-circulated SPSS package. Use of this routine is straightforward: the user must simply identify the finishing state variable (coded zero for event observed and unity for censored cases) and the finishing time variable (starting time is assumed zero). The routine returns numerical estimates of the survivor function but also graphical displays such as Figure 8, which shows the estimated survivor function for job changes as a function of the duration in the job using Mayer's data.

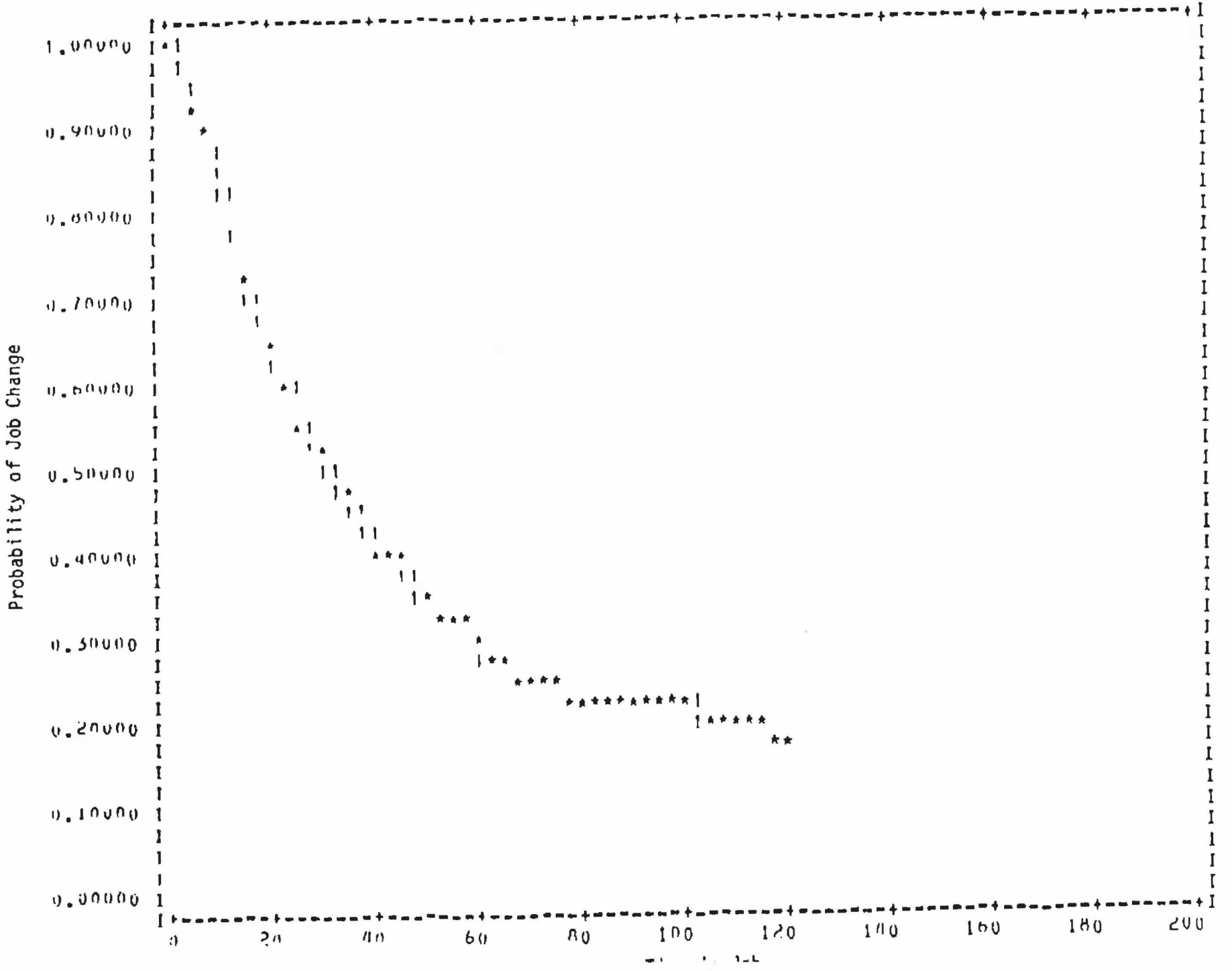
Figure 8 About Here

Figure 8 is straightforward to interpret. The vertical axis gives the probability of no job change - 'surviving' in the current job. The horizontal axis gives the length of time in the job measured in months. The plotted points show the Kaplan-Meier estimates: the numeral one is plotted for single points and the asterisk for one or more points grouped closely together. The estimated survival curve shows that the probability of staying in a job for the first twenty months is high: approximately .70. By the sixtieth month, however, only 25% of these persons remain the their jobs. The probability of remaining in the job after this point levels off considerably so that by 120 months approximately 18% still have not changed jobs.

GRAPH OF SURVIVAL FUNCTION

SURVIVAL VARIABLE V13

Figure 8. Survivor Plot of Job Change Data



Survivor plots similar to Figure 8 are useful descriptive tools but they do not aid much in model searches. For this task, we turn to the log survivor plot, which is simply a plot of G(t) against time, or G(t) plotted on a log scale. The log survivor plot is useful for model searches because, as we saw much earlier, the plot will be linear when the transition rate is time-independent. When the log-survivor plot is nonlinear, it suggests a more complex model. The log survivor plot is also useful because the negation of its slope yields an empirical estimate of the transition rate.

Figure 9 gives the log survivor plot for the rate job changes using Mayer's data. The plotted points are more linear than those in Figure 8 but they do not resemble a straight line. More accurately, the plot resembles a smooth curve. This observation suggests that the rate is not time-independent and that the investigator may wish to consider a model with time-dependent transition rates. In addition, the decreasing negative slope of the plot suggests a model where the rate is a declining function of the time in the job.

Figure 9 About Here

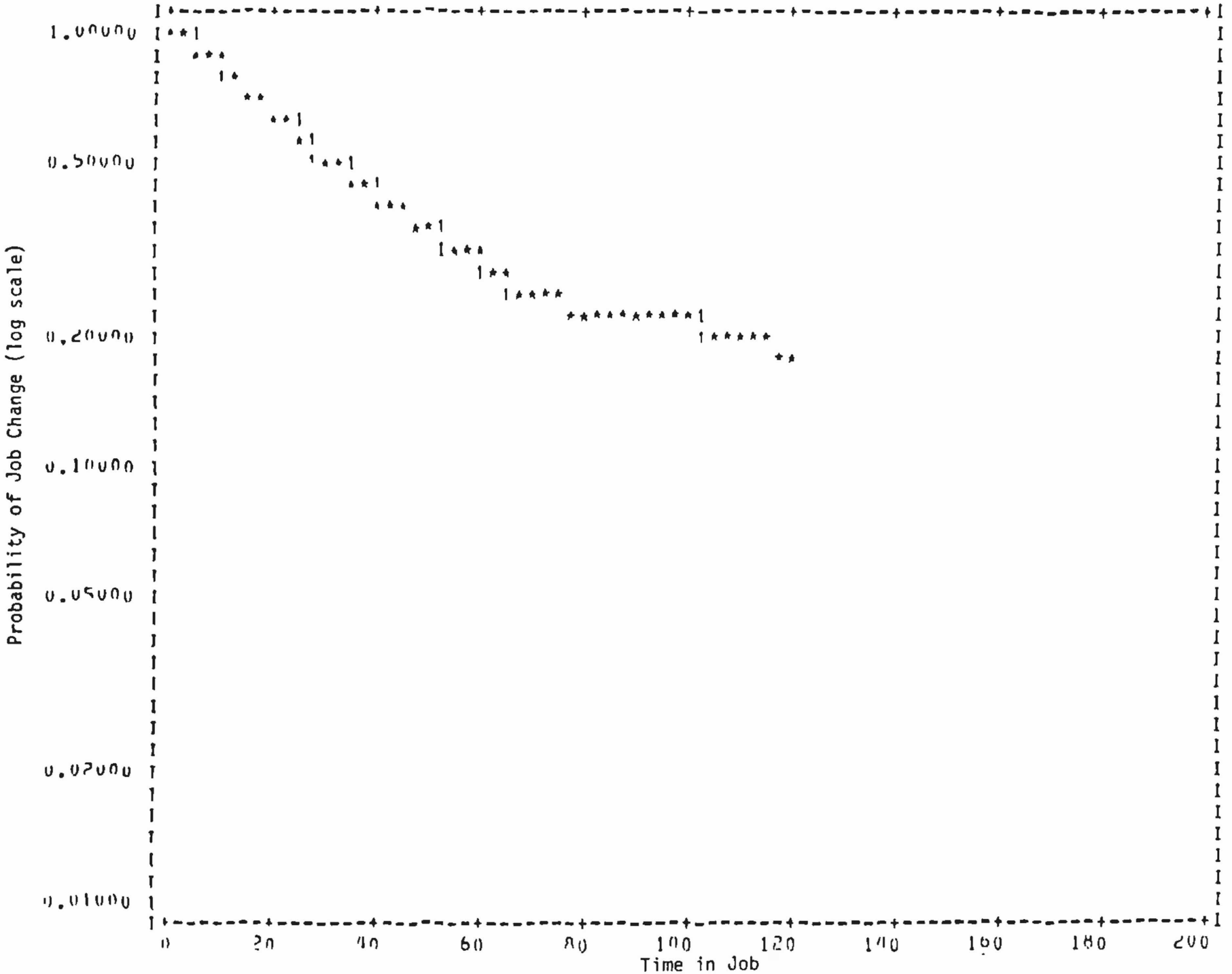
Time-dependence in the transition rates is one interpretation for non-linear log survivor plots; another, equally plausible, interpretation is population heterogeneity. That is, if the data contain several subgroups, each characterized by time-independent rates of different values, then the log survivor plot might be nonlinear. Unfortunately, there is no definitive method for choosing between these alternative explanations. Instead, the investigator must explore the data for heterogeneity and then finally make a decision on time-dependence in light of substantive considerations.

This issue can be investigated with the data job changes. Our position is that we have evidence of time-dependence but wish to explore the possibility

LOG SCALE GRAPH OF SURVIVAL FUNCTION

SURVIVAL VARIABLE V13

Figure 9. Log Survivor Plot of Job Change Data



that it was generated by heterogeneity. The strategy is to divide the data into subgroups according to variables such as sex which are thought to affect the rate of job change. If we continue to see nonlinear log survivor plots across these subgroups, then our confidence in the time-dependence interpretation increases. This strategy has the additional advantage of bringing important independent variables to our attention and allowing a nonparametric comparison of their effects on the transition rates.

Figures 10 and 11 show examples of this strategy. Figure 10 presents the log survivor plot of job changes for subgroups of males and females. The plot for females uses the numeral 2 and the male plot uses 1. Both plots are nonlinear but their slopes vary, especially in the first twenty months. These observations suggest that the nonlinear character of the aggregate plot (Figure 9) is not due to sexual heterogeneity although sex differences apparently play a role in the job change process.

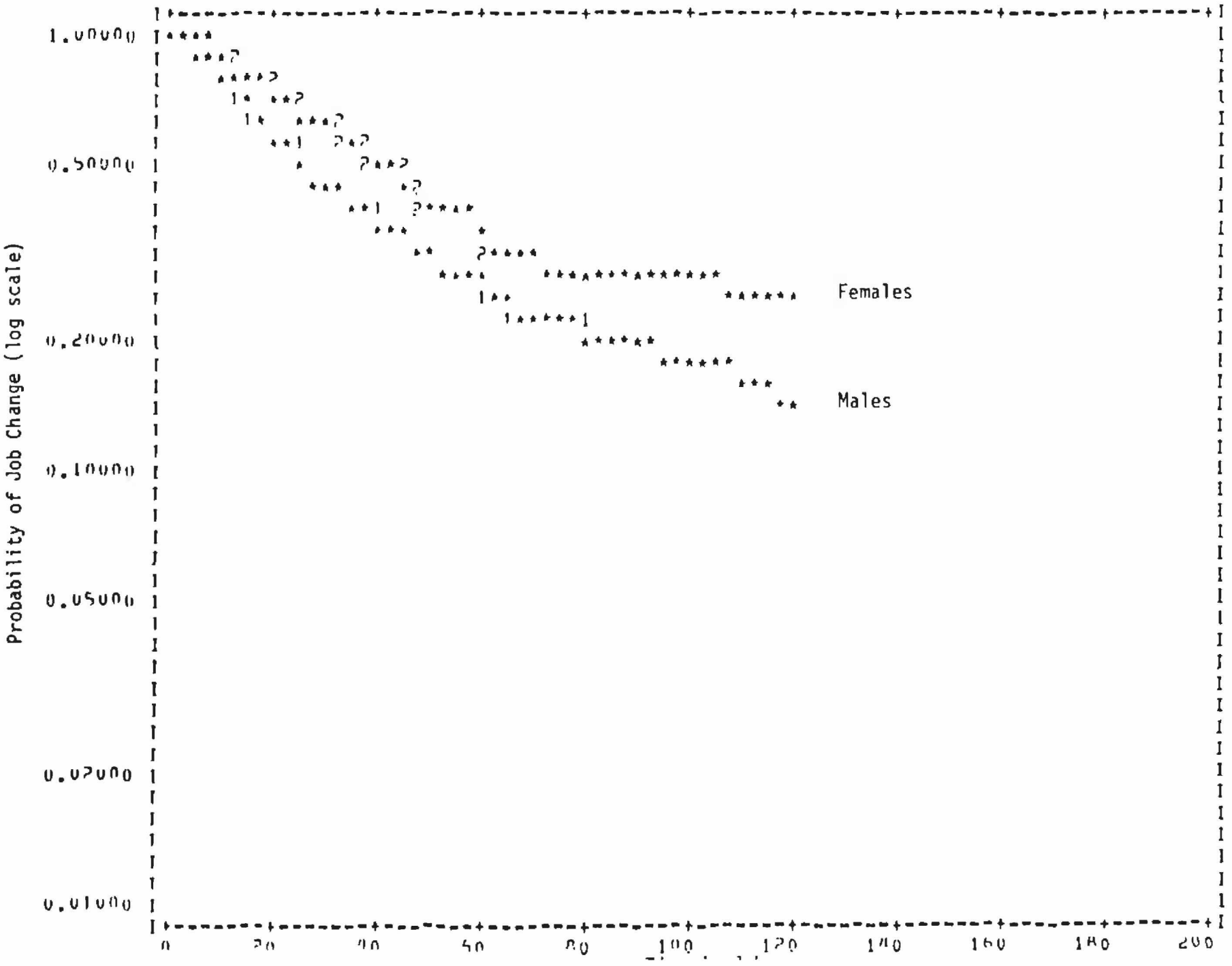
Figures 10 and 11 About Here

Figure 11 shows a similar set of plots for subgroups of private-firm (numeral 1) and public-firm (numeral 2) jobs. The interpretation of this figure is again the same: we see evidence of differences in the rate for the different types of firms but jobs in both types of firms continue to show the apparent time-dependence observed in the aggregate plot. This additional finding begins to throw the weight of the evidence on the time-dependent interpretation and against the heterogeneity interpretation. However, in a substantive research application, we would probably want to examine many more subgroups before we made a final modeling decision.

So far in the exploratory analysis of job changes, we have examined only the survivor function for job changes of all variety. Depending on substantive

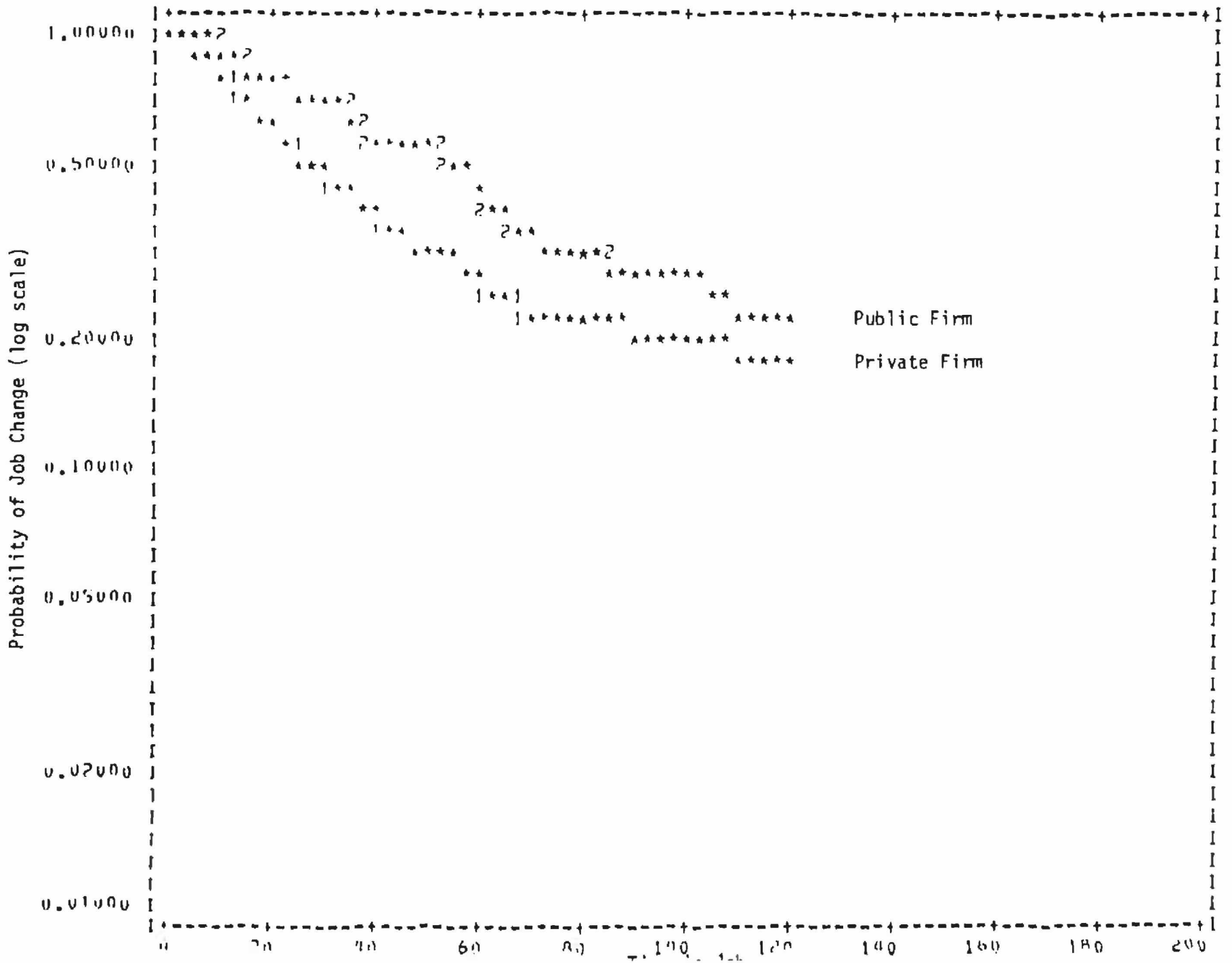
SURVIVAL VARIABLE V13
GROUPED BY V1

Figure 10. Log Survivor Plot for Job Changes by Sex



SURVIVAL VARIABLE V13
GROUPED BY V55

Figure 11. Log Survivor Plot for Job Changes by Type of Firm



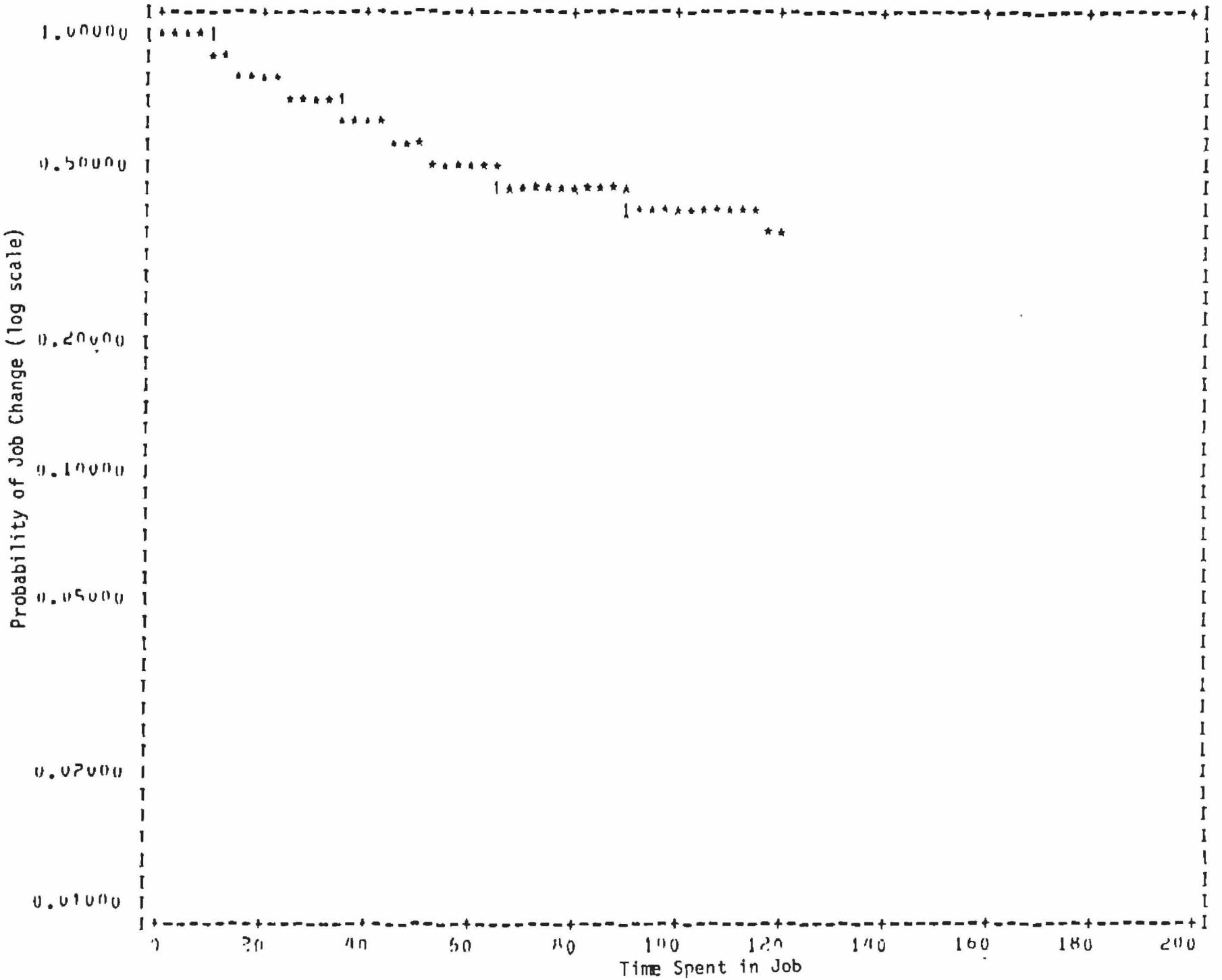
considerations, the investigator may also wish to explore the survivor functions for specific types of job changes only. This activity requires a shift from the aggregate state change variable to a more disaggregated one - such as those discussed in the data structure section. For instance in the job change data, an investigator might be curious about differences in job changes that are voluntarily initiated by the employee and those that are initiated by the employer. This examination cannot be conducted by the use of independent variables since they will be confounded by the censoring problem - only those cases for which a change has been observed will it be possible to measure these variables accurately. Consequently, the investigator must use a disaggregated destination state variable or else separate destination state variables for each outcome.

Figure 12 shows the survivor plot for the voluntary job changes and Figure 13 shows a similar plot for firm-initiated job changes. The plot for voluntary changes looks similar to the earlier plots except that it has a gentler slope, reflecting the fewer number of events. The plot for firm-initiated changes is remarkably different: it is nearly linear and it has only a slight slope. These observations might prove of great interest to the investigator for they clearly suggest that different processes generate the two types of job changes. Moreover, they suggest that the models to be used in further analysis should be quite different: for voluntary changes, a time-dependent model appears appropriate; for firm-initiated changes, a time-independent model appears best. Of course, before proceeding to such models, the investigator will wish to examine subgroups for each of these plots.

Figures 12 and 13 About Here

SURVIVAL VARIABLE VTS

Figure 12. Log Survivor Plot for Voluntary Job Changes



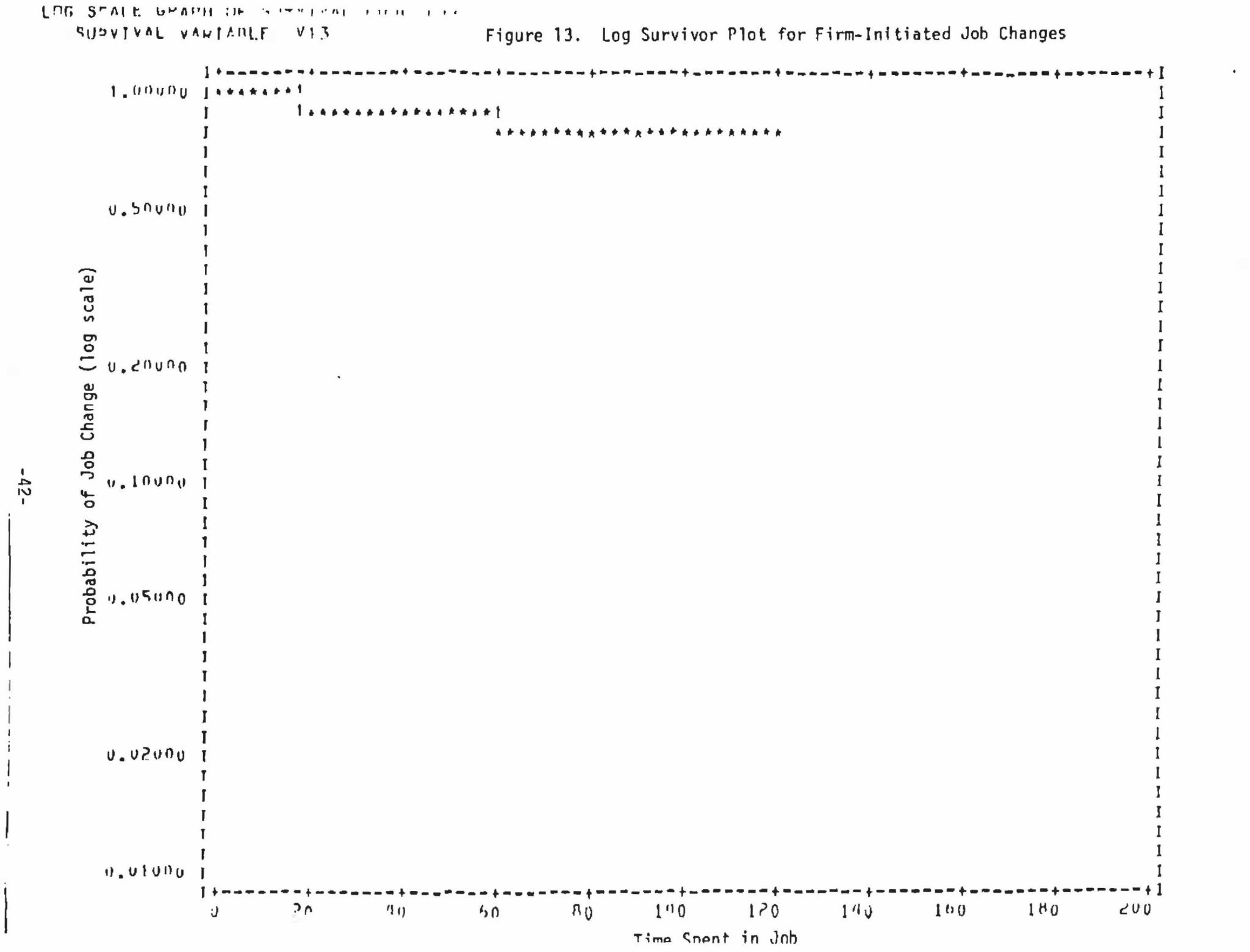


Figure 13. Log Survivor Plot for Firm-Initiated Job Changes

5. Model Specification and Estimation

Log survivor plots are useful exploratory tools but they can also be used for hypothesis testing. There are several statistical tests available for comparing survival curves. Nonetheless, in most sociological applications, the comparisons of interest will either be so numerous as to make this strategy intractable or else will involve subgroups of the data so small as to make the statistical tests meaningless. When the common multivariate analysis is desired, it is more feasible to specify a model of the process and to estimate the parameters of this model. Unlike survival analysis, this approach has the advantage of describing completely the form of the relationships between independent variables and the rate of transition.

Similar to regression analysis, where a linear specification is frequently chosen for simplicity, multivariate analysis of rates is commonly applied with a simple baseline model. Initially, this baseline model was also a linear model of the form

$$r(t) = \alpha_0 + \alpha_1 X_1(t) + \cdots + \alpha_m X_m(t) \tag{27}$$

where the X(t)'s are the independent variables and the parameters measure their effects on the rate. However, since the linear model can lead to negative predicted rate (a theoretical impossibility), the accepted baseline model has come to be the log-linear specification

$$\ln\ r\ (t)\ = \alpha_0 + \alpha_1 X_1(t) + \cdots + \alpha_m X_m(t) \tag{28}$$

which constrains all predicted values to be positive.

The log linear model can be rewritten several equivalent ways:

$$r(t) = \exp(\alpha_0 + \alpha_1 X_1(t) + \cdots + \alpha_m X_m(t)) \quad (29.1)$$

$$r(t) = e^{\alpha_0} e^{\alpha_1 X_1(t)} \cdots e^{\alpha_m X_m(t)} \quad (29.2)$$

$$r(t) = a_0 a_1^{X_1(t)} \cdots a_m^{X_m(t)} \quad (29.3)$$

and the parameters $a_m = e^{\alpha_m}$ are intuitively understood as the multipliers of the base rate a_0. It is critical to recognize here that a small value of say α_1 might translate to an a_1 parameter of 1.05. Since the variable $X_1(t)$ is the order of this parameter, when it takes a wide range of values enormous differences in the predicted rate will result. One way to obtain an intuitive feel for the magnitude of these effects is to think of $100(a_m - 1)$ as the percentage increase or decrease in the rate due to each unit increment in the variable $X(t)$.

Although the log linear specification is commonly assumed by default, the investigator will sometimes have reason to specify a more complex model of the process. The motivation for this specification might come from an exploratory survival analysis, previous research, or a developed theory. For example, our explorations with Mayer's data leads us to expect that the rate of job change will decline with time in the job. A model which has been used previously for this process is known as the Gompertz model and assumes the rate declines exponentially with time. It is specified by the equation

$$r(t) = e^{\alpha_0} e^{\xi_0 t}, \quad \xi_0 < 0 \quad (30)$$

where the parameter ξ_0 measures the time-dependence in the process and is expected to be negative. The corresponding survivor function for the

Gompertz model is

$$G(t) = \exp\{(-\frac{e^{\alpha_0}}{\beta_0})\ (e^{\beta_0 t} - 1)\} \tag{31}$$

and shows the way complexity in the rate reverberates through the model. Other common specifications for the functional form of the rate are presented in Appendix C.

Our exploratory analysis with the job change data showed not only time-dependence but also population heterogeneity. Consequently, we want to introduce independent variables into the Gompertz model. This can be achieved in either of two ways. First, the variables can be included in the time-independent vector as

$$r(t) = \exp\ (\alpha_0 + \alpha_1 X_1(t))\ e^{\beta_c t} \tag{32}$$

Second, they can be specified in interaction with the time-dependence:

$$r(t) = e^{\alpha_0} \exp\ (\beta_0 + \beta_1 X_1(t))t \tag{33}$$

These two models are fundamentally different: in the first, time-dependence is merely controlled; in the second, the effect of a variable depends not only on its value but also on time. It is, of course, also possible to specify a combined model with variables in both vectors, however, extreme caution must be exercised with the time-dependent vector. When the overall prediction for the time-dependent vector is positive, the process reverses from one in which the rate declines exponentially with time to one in which the rate increases exponentially with time.

In many applications, the investigator will have no advance reason for preferring one Gompertz specification over the other and it is probably

best to assume the simpler, time-independent specification of heterogeneity. Moreover, when the exploratory analysis shows, as with the job change data, that despite group differences, the form of time-dependence appears similar, this specification can be used with confidence. In later analysis, the investigator may wish to compare this model against the more complex one.

After the researcher has chosen a model, or a set of models to compare, the task is to estimate the parameters. In the simplest case, where a constant rate model is preferred and the data are uncensored, ordinary least squares regression might be used. This approach takes advantage of the expected time until a change of state which for this model is simply

$$E(t) = \frac{1}{r} = \frac{1}{e^{a_0 + a_1 X_1}} = e^{-a_0} e^{-\alpha_1 X_1} \tag{34}$$

This can be transformed to the log-linear equation

$$E(\ln t) = \int_0^\infty \ln t \exp(-rt)dt \tag{35}$$

which reduces to

$$\begin{aligned} E(\ln t) &= \text{Euler's constant} - \ln r \\ &= .577217 - \ln r \\ &= .577217 - a_0 - a_1 X_1 \end{aligned} \tag{36}$$

Since this equation is linear, it can be estimated by least squares using ln(t) as the dependent variable. The regression estimates can then be be transformed (reversed) to recover the parameters α of the model of interest.

The regression strategy is useful because it requires no special tools, however, its use requires very restrictive conditions which are uncommon in

the social sciences. Most social science data contain censored cases and the constant rate model is often inappropriate. In either instance, the least squares strategy will lead to biased estimates of the model and an alternative strategy must be used. The approach which I will emphasize here is maximum likelihood estimation and I will discuss it in the context of Nancy Tuma's computer program RATE.

A complete understanding of maximum likelihood estimation is beyond the scope of this paper. Nonetheless, I shall try to sketch an outline of the procedure as it applies to rate models with event-history data and only right censoring. The general strategy of maximum likelihood estimation involves first writing a likelihood function L for the joint probability of the observations. This joint probability is simply the product of the individual probabilities L_i for each observation on case i. Recall that each individual observation contains two pieces of information: the state occupied and the time in the state. Thus,

$$L_i = \Pr\left[\overline{Y = y, T = t}\right] \tag{37}$$

for any case i. If we consider only single, irreversible events, then we have, in addition to the time information, two types of state observations: those units for which the event is observed (y = 1) and those for which it is censored (y = 0). Now let $\emptyset$ be an indicator variable that takes the value of unity for uncensored cases and zero for censored cases. The likelihood for case i can now be divided into two components

$$L_i = \Pr\left[\overline{Y = 0, T = \underline{t}}^{1-\emptyset}\right] \cdot \Pr\left[\overline{Y = 1, T = \underline{t}}^{\emptyset}\right] \tag{38}$$

the first which gives the probability of no event by time t and the second which gives the probability of an event at exactly time t. However, we know

from above that the probability of no event by t is the survivor function G(t). The probability of an event at exactly t is given by the probability density function f(t), but this function can be rewritten as the product of the rate and the survival function. (In other words, the probability of an event at t is the product of survival until t and the rate of events at t, see Appendix C.) Using these substitutions, we can write

$$L_i = G_i(t)^{1-\emptyset} \left[r_i(t).G_i(t)\right]^{\emptyset} \qquad (39)$$

which can be reorganized as

$$L_i = G_i(t).r_i(t)^{\emptyset} \quad . \qquad (40)$$

Specific models of rates are estimated by substituting the appropriate equations in the likelihood function and finding the parameters which maximize it. In other words, we wish to find the parameters which predict best the likelihood of the sample observations. In practice, finding this maximum is numerically tedious and one must usually search iteratively for the best parameters using a computer routine. The best available program for the models we are considering is Tuma's RATE, however, others are available (including Coleman's programs and partial likelihood procedures in SAS and BMDP). Tuma's program has the advantage of estimating a wide-range of models for several types of data structures. It also returns the standard errors of the parameter estimates of each model.

Table 2 presents a comparison of least squares and maximum likelihood estimators for models with a single covariate using Mayer's job change data. The top horizontal line of the table gives three separate sets of estimates for the effect of the dummy variable for public firms on the rate of job change. The first estimates were obtained with least squares regression

using the log of time as the dependent variable and no constraints. The second set of estimates were given by RATE when a time-independent model was specified. The third group of estimates are also from RATE but from a Gompertz specification with the covariate in the time-independent vector.

Table 2 About Here

All three sets of estimates for the public firm variable are negative, suggesting that the rate of job change is lower in the public sector. The three sets of estimates are also remarkably similar, however, the significant estimate of the time-dependent parameter β_o suggests that the Gompertz model is to be preferred. The second line of the table shows the estimates for the effect of age on the rate of job change. Again, the different estimates agree generally - all show small negative effects - although the least squares estimate is smallest.

The final two lines of the table report estimates for the same models but using data on first jobs only. In theory, these data should yield closer agreement between the least squares and maximum likelihood estimates since the level of censoring is lower. Table 2 shows that in practice this is not always true. The age variable parameter estimates do show closer agreement than with all data but the estimates for the public firm variables do not - in fact, they diverge considerably. These comparisons may be irrelevant, however, since the estimates are all statistically insignificant. This suggests that a substantive difference with the first jobs has contaminated our comparison.

In most research applications, the investigator will wish to estimate models with more than one exogenous variable. For example, the rate of job change is thought to be a function of personal characteristics such as age,

Table 2. Comparison of Estimates Across Estimators and Models (Standard errors shown in parentheses)

Variables included in the Model	Least Squares Estimates of Constant Rate Model		Maximum Likelihood Estimates of Constant Rate Model		Maximum Likelihood Estimates of Time-dependent Gompertz Model		
	$\hat{\alpha}_o$	$\hat{\alpha}_1$	$\hat{\alpha}_o$	$\hat{\alpha}_1$	$\hat{\alpha}_o$	$\hat{\alpha}_1$	$\hat{\beta}_o$
Dummy variable for public firms with all data (22.2% censoring)	-3.22	-.282 (.145)	-4.02 (.058)	-.388 (.158)	-3.66 (.015)	-.402 (.158)	-.008 (.001)
Age in months at start of job with all data (22.2% censoring)	-2.79	-.002 (.001)	-2.65 (.211)	-.005 (.001)	-2.34 (.215)	-.005 (.001)	-.008 (.001)
Dummy variable for public firms with data on first jobs only (8.1% censoring)	-3.02	-.584 (.260)	-3.61 (.113)	-.407 (.300)	-3.25 (.141)	-.443 (.300)	-.010 (.003)
Age in months at start of job with data on first jobs only (8.1% censoring)	-2.72	-.002 (.003)	-3.29 (.470)	-.002 (.002)	-2.78 (.550)	-.002 (.002)	-.010 (.003)

sex, education and experience; of job characteristics such as wage, autonomy and hierarchiacal position; and of organizational characteristics such as size and sector. Table 3 reports the effects of these variables in three different equations. The top equation estimates only the constants of the Gompertz model, the second equation includes the personal characteristics and wage, and the third equation adds the organizational and positional variables.

Table 3 About Here

The estimates in Table 3 show that all the variables except education have negative effects on the rate of changing jobs. Inspection of the standard errors show that most of these effects are statistically significant, with the exception of SUBS, WAGE and the COH variables.

A second method for evaluating the importance of variables in a model is to compare the relative fits of nested hierarchical models. This test, known as the likelihood ratio test, uses the knowledge that minus two times the log of the ratio of the likelihoods of a constrained model over an unconstrained one is Chi square distributed. That is, if L_Ω is the likelihood for a model with q+s parameters and L_w is the likelihood for a nested model with only q free parameters (and s constrained parameters), then when $\lambda = \max L_w / \max L_\Omega$ the value $-2 \ln \lambda$ is χ^2 distributed with s degrees of freedom.

Table 3 shows the Chi square values and the degrees of freedom for each of the three models compared to a constant rate model with no exogenous variables. For the first equation, we see that since a χ^2 value of 37.9 with 1 degree of freedom is significant, the inclusion of the time-dependent constant significantly improves the model. The values of the χ^2 statistic

Table 3. Maximum Likelihood Estimates of Effects of Organizational and Persoanl Characteristics on Rate of Job-Shifts (Standard Errors shown in parentheses)

Dependent Variable	Independent Variables													
	Constant	WFX	EDUC	FEMALE	COH 31	COH 41	WAGE	SIZE	PUBLIC	SUBS	JAUTO	Time dependent constant	$2 \times LR$	DF
Move	-3.74 (.091)											-.009 (.002)	37.9	1
Move	-3.09 (.384)	-.005 (.001)	.0003 (.002)	-.563 (.163)	-.252 (.226)	-.060 (.192)	-.0001 (.0002)					-.008 (.002)	77.3	7
Move	-2.81 (.416)	-.003 (.002)	.006 (.003)	-.476 (.168)	-.222 (.321)	-.085 (.194)	-.0001 (.0004)	-.067 (.028)	-.579 (.218)	-.004 (.009)	-.259 (.057)	-.006 (.002)	108.	11

Variable Definitions

MOVE = rate of leaving a job

WFX = work force experience in months

EDUC = total education in months

FEMALE = dummy for female

COH 31 = dummy for 1931 cohort member

COH 41 = dummy for 1941 cohort member

WAGE = monthly wage in DM

SIZE = number of employees in organization

PUBLIC = dummy for public sector firm

SUBS = number of subordinates

JAUTO = scale of job autonomy

for the other two equations show that they also are improvements relative to the homogenous constant rate.

The models in Table 3 can also be compared relative to each other since they are nested hierarchically. The test here is simply the difference of the Chi square values evaluated by the difference in degrees of freedom. For example, to test the fit of the second equation relative to the first, we find the Chi square value 77.3 - 37.9 = 39.4 and use the degrees of freedom 7 - 1 = 6. Looking this value up in any standard χ^2 table shows that it is significant and thus the second model is a significant improvement over the first. A similar test for the third equation shows that it is also an improvement over either of the two nested models.

As in the survival analysis, the investigator may wish to disaggregate the model into different destination states. In the job change analysis, I wanted to see if the organizational variables showed different effects depending on the relative wage of the next job. Consequently, I disaggregated the simple job change variable into a variable with 3 types of job changes or destination states. The first state is for movement to jobs with 15% or greater wage increase and indicates upward mobility. The second state is for movement to lower paying jobs. The third state is for "lateral" moves to jobs with an increase of less than 15% but no decrease. I then reestimated the model with and without the organizational variables for each destination state. Table 4 presents the findings.

Table 4 About Here

Applying the likelihood ratio test to the nested models in Table 4, we find that the organizational variables are important for the upward and

Table 4. Maximum Likelihood Estimates of Effects of Organizational and Personal Characteristics on Rate of Directional Job-Shifts (Standard Errors shown in parentheses)

Dependent Variable	Independent Variables													
	Constant	WFX	EDUC	FEMALE	COH 31	COH 41	WAGE	SIZE	PUBLIC	SUBS	JAUTO	Time dependent constant	χ^2_{LR}	DF
Upward	-3.96 .625)	-.004 (.002)	.001 (.004)	-.622 (.247)	.089 (.396)	.634 (.344)	-.001 (.000)					-.010	61.3	7
Upward	-3.01 (.666)	-.001 (.002)	.007 (.004)	-.560 (.257)	.031 (.411)	.625 (.351)	-.001 (.000)	-.092 (.044)	-.061 (.318)	-.011 (.016)	-.439 (.098)	-.007 (.003)	86.3	11
Down	-6.39 (.888)	-.009 (.004)	.003 (.005)	.114 (.376)	.678 (.519)	.147 (.473)	.001 (.000)					-.007 (.004)	15.2	7
Down	-6.61 1.04)	-.009 (.004)	.015 (.006)	.211 (.392)	.780 (.518)	.103 (.474	.001 (.000)	-.167 (.071)	-2.11 (.794)	.011 (.014)	-.275 (.124)	-.004 (.004)	36.7	11
Lateral	-3.82 (.679)	-.004 (.003)	-.0004 (.004)	-1.00 (.318)	-.963 (.394)	-.666 (.323)	-.0001 (.0003)					-.005	34.2	7
Lateral	-4.04 (.764)	-.003 (.003)	.002 (.005)	-.895 (.326)	.984 (.405)	-.695 (.328)	-.0001 (.0003)	-.036 (.048)	-.377 (.359)	-.018 (.032)	-.060 (.102)	-.004 (.003)	37.1	11

Variable Definitions

Upward = rate of movement to jobs with 15% or greater wage increase

Down = rate of movement to jobs with wage decrease

Lateral = rate of movement to jobs with wage increase less than 15% but no decrease

For other variables, see Table 3.

downward mobility processes but not for the lateral process. Within the upward process, we see that size and job autonomy both decrease the rate of movement; the other organizational variables are insignificant. For downward mobility, however, the public firm variable also shows a large negative effect. Thus, being in a public firm does not improve one's chances of getting a higher paying job but does decrease dramatically the possibility that one will move to a lower paying job. Findings of this variety are impossible to see from Table 3, which simply shows a negative effect of public firms on all movement and leaves a quite different image of the process.

6. Concluding Remarks

I have tried in this paper to give an introduction to the dynamic analysis of discrete dependent variables. I have sketched only the outlines of the basic methodological approach. It is my intention that this outline should provide the reader with sufficient ammunition to tackle the methodology in greater detail and with more sophistication. For those with whom this intention has been fulfilled, I have provided a list of additional readings. in Appendix D.

Appendix A

Definitions of Mathematical Terms

State space variable — $Y_n(t)$, integer-valued for n events

State probability — $\Pr\left[Y(t) = y\right]$

Transition probability — $q_{jk}(t,t+\Delta t) = \Pr\left[Y(t+\Delta t) = k \mid y(t) = j\right]$

Conditional transition probability — $m_{jk}(t) = \Pr\left[Y_n = k \mid Y_{n-1} = j\right]$

Instantaneous transition rate — $r_{jk}(t) = \lim_{\Delta t \to o} \frac{q_{jk}(t,t+\Delta t)}{\Delta t}$

Hazard function — $h_j(t) = \sum_{\substack{k \\ k \neq j}} r_{jk}(t)$

Cumulative hazard function — $H_j(t) = \int_o^t h_j(s)ds$

Survivor function — $G_j(t) = \Pr\left[T_n \geq t\right]$

Cumulative distribution function — $F_j(t) = \Pr\left[T_n \leq t\right]$

Probability density function — $f_j(t) = \lim_{\Delta t > o} \frac{\Pr\left[t < T < t+\Delta t\right]}{\Delta t}$

Appendix B

Useful Relationships Between Terms

$$h_j(t) = r_j(t)$$

$$r_j(t) = \sum_{\substack{k \\ k \neq j}} r_{jk}(t)$$

$$r_{jk} = r_j(t).m_{jk}(t)$$

$$r_j(t) = \frac{-d \ln G_j(t)}{dt}$$

$$r_j(t) = f_j(t)/G_j(t)$$

$$m_{jk}(t) = r_{jk}(t)/r_j(t)$$

$$G_j(t) = 1-F_j(t)$$

$$G_j(t) = \exp\left[-\int_0^t r_j(s)ds\right]$$

$$F_j(t) = 1-G_j(t)$$

$$F_j(t) = \int_0^\infty f(t)\, dt$$

$$f_j(t) = \frac{d\, F_j(t)}{dt}$$

$$f_j(t) = \frac{-d\, G_j(t)}{dt}$$

$$f_j(t) = r_j(t).G_j(t)$$

$$H_j(t) = -\ln G_j(t)$$

$$E(t) = \int_0^\infty t\, f_j(t)\, dt$$

$$Var(t) = \int_0^\infty t^2 f(t)dt - E(t)$$

$$E(N_t) = \int_0^t r_j(s)\, ds$$

Appendix C

Common Functional Forms of the Transition Rate

Constant rate model	$r_{jk}(t) = \alpha$
Rayleigh	$r_{jk}(t) = \alpha + 2\beta t$
Weibull	$r_{jk}(t) = \alpha\beta t^{\beta-1}$
Gompertz	$r_{jk} = \beta\exp(\gamma t)$
Makeham's Law	$r_{jk}(t) = \alpha + \beta\exp(\gamma t)$
Double exponential	$r_{jk}(t) = \alpha\beta\exp(-\beta t) / \{1-\alpha[1-\exp(-\beta t)]\}$
Gamma	$r_{jk}(t) = \dfrac{t^{\gamma-1}\exp(-\gamma t)}{\int_t^\infty X^{\gamma-1}\exp(-\lambda X)dX}$

Appendix D

Additional Reading

Bartholomew, D. J. 1973. Stochastic Models for Social Processes. 2nd ed. New York: Wiley.

Breslow, N. 1974. "Covariance analysis of censored data," Biometrika 30: 89-99.

Carroll, G. R. and J. Delacroix. Forthcoming. "Organizational mortality in the newspaper industries of Argentina and Ireland: an ecological approach," Administrative Science Quarterly.

Coleman, J. S. 1964. Introduction to Mathematical Sociology. Glencoe, Ill.: Free Press.

_____1968. "The mathematical study of change," Pp. 428-78 in Methodology in Social Research. edited by H. Blalock and A. Blalock. New York: McGraw-Hill.

_____1973. The Mathematics of Collective Action. Chicago: Aldine.

_____Forthcoming. Longitudinal Data Analysis

Cox, D. R. 1972 "Regression models and life tables," Journal of the Royal Statistical Society, Series B, 34: 187-220.

_____1975. "Partial likelihood," Biometrika 62: 269-276.

Cox, D. R. and P. A. W. Lewis. 1966. The Statistical Analysis of Series of Events. London: Methuen.

DiPrete, T.1981. "Unemployment over the life-cycle:racial differences and the effects of changing economic conditions," American Journal of Sociology 87: 286-307.

Elandt-Johnson, R. C. and N. L. Johnson. 1980. Survival Models and Data Analysis. New York: Wiley.

Flinn, C. J. and J. J. Heckman. 1980. "Models for the analysis of labor force dynamics," Discussion paper #80-3, Economics Research Center, National Opinion Research Center. Chicago, Illinois.

Cross, J. J. and V. A. Clark. 1975. Survival Distributions: Reliability Applications in the Biomedical Sciences. New York: Wiley.

Hannan, M. T. and G. R. Carroll. 1981. "Dynamics of formal political structure: an event-history analysis," American Sociological Review 46: 19-35.

Hannan, M. T., N. B. Tuma and L. P. Groeneveld. 1977. "Income and marital events: evidence from an income maintenance experiment," American Journal of Sociology 82: 1186-1211.

_____. 1978 "Income and independence effects on marital dissolution: results from the Seattle and Denver Income Maintenance Experiments," American Journal of Sociology 84: 611-633.

Kalbfleisch, J. D. and R. L. Prentice. 1981. The Statistical Analysis of Failure Time Data. New York: Wiley.

Kaplan, E. L. and P. Meier. 1958. "Nonparametric estimation from incomplete observations," Journal of the American Statistical Association 53: 457-481.

Sheps, M. C. and J. A. Menken. 1972. Mathematical Models of Conception and Birth. Chicago: University of Chicago Press.

Singer, B. and S. Spilerman. 1974. "Social mobility models for heterogeneous populations," Pp. 256-401 in Sociological Methodology 1973-74, edited by H. Costner. San Francisco: Jossey-Bass.

_____.1976. "The representation of social processes by Markov models," American Journal of Sociology 82: 1-54.

Sørensen, A. B. and N. B. Tuma. 1978. "Labor market structures and job mobility," Institute of Research on Poverty, Working Paper #505-78, University of Wisconsin, Madison.

Spilerman, S. 1972. "The analysis of mobility processes by the introduction of independent variables into a Markov chain," American Sociological Review 37: 277-294.

Tuma, N. B. 1976. "Rewards, resources and the rate of mobility: a nonstationary multivariate stochastic model," American Sociological Review 41: 338-360.

Tuma, N. B. and M. T. Hannan. 1979. "Approaches to the censoring problem in analysis of event histories," In Sociological Methodology 1979, edited by K. Schuessler. San Francisco: Jossey-Bass.

_____. 1979. "Dynamic analysis of event histories," American Journal of Sociology 84: 820-854.